ABSTRACT

Disclosed are a computer-readable code, system and method for classifying a target document in the form of a digitally encoded natural-language text as belonging to one or more of two or more different classes. For each of a

5   plurality of non-generic words and/or words groups characterizing the target document, there is determined a selectivity value calculated as the frequency of occurrence of that term in a library of texts in one field, relative to the frequency of occurrence of the same term in one or more other libraries of texts in one or more other fields, respectively, and the document is represented as a vector of terms,

10   where the coefficient assigned to each term is a function of the selectivity value determined for that term. There is then determined, for each of the plurality of sample texts having associated classification identifiers, a match score related to the number of descriptive terms present in or derived from that text that match those in the target text. From the selected matched texts, and the associated

15   classification identifiers, a classification determination of the target document is made.